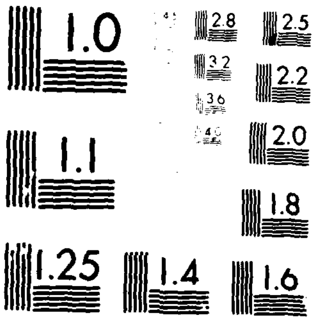
AD-A084 262   MINNESOTA UNIV ST PAUL DEPT OF APPLIED STATISTICS   F/G 12/1
CONTINGENCY TABLES.(U)
FEB 80   S E FIENBERG   N00014-78-C-0600
UNCLASSIFIED   TR-369   NL

1 OF 1
AD
AD A084262

END
DATE
FILMED
6-80
DTIC


1.0
2.8
2.5
3.2
2.2
3.6
4.0
2.0
1.1
1.8
1.25
1.4
1.6

UNIVERSITY OF MINNESOTA

SCHOOL OF STATISTICS

2

CONTINGENCY TABLES*

by

Stephen E. Fienberg

Technical Report No. 369

School of Statistics

University of Minnesota

February, 1980

*This is a draft of an article prepared for the Encyclopedia of Statistical Sciences, to be published by John Wiley & Sons, Inc.

1. Introduction and Historical Remarks

Multivariate statistical analysis has occupied a prominent place in the classical development of statistical theory and methodology. The analysis of cross-classified categorical data, or contingency table analysis as it is often referred to, represents the discrete multivariate analogue of analysis of variance for continuous response variables, and now plays an important role in statistical practice. This presentation is intended as an introduction to some of the more widely used techniques for the analysis of contingency table data, and to the statistical theory that underlies them.

The term contingency, used in connection with tables of cross-classified categorical data seems to have originated with Karl Pearson [1904], who for an s×t-fold table defined contingency to be any measure of the total deviation from "independent probability" The term is now used to refer to the table of counts itself. Prior to this formal use of the term, statisticians going back at least to Quetelet [1849], worked with cross-classifications of counts to summarize the association between variables. Pearson [1900a] has laid the groundwork for his approach to contingency tables, when he developed his χ^2 test for comparing observed and expected (theoretical) frequencies. Yet Pearson preferred to view contingency tables involving the cross-classification of two or more polytomies as arising from a partition of a set of multivariate, normal data, with an underlying continuum for each polytomy. This view led Pearson [1900b] to develop his tetrachoric correlation coefficient for 2×2 tables, and this work in turn spawned an extensive literature well chronicled by Lancaster [1969].

The most serious problems with Pearson's approach were (1) the complicated infinite series linking the tetrachoric correlation coefficient with the frequencies in a 2×2 table and (2) his insistence that it always made sense to assume an underlying continuum, even when the dichotomy of interest was dead-alive or employed-unemployed, and that it was reasonable to assume that the probability distribution over such a continuum was normal. In contradistinction, Yule [1900] chose to view the categories of a cross-classification as fixed, and he set out to consider the structural relationship between or among the discrete variables represented by the cross-classification, via various functions of the cross-product ratio. Especially impressive in this, Yule's first paper on the topic, is his notational structure for n attributes or 2^n tables, and his attention to the concept of partial and joint association of dichotomous variables.

The debate between Pearson and Yule over whose approach was more appropriate for contingency table analysis raged for many years (see e.g., Pearson and Heron [1913]), and the acrimony it engendered was exceeded only by that associated with Pearson's dispute with R.A. Fisher over the adjustment in the degrees of freedom (d.f.) for the chi-square test of independence in the s×t-fold table. (In this latter case Pearson was simply incorrect; as Fisher [1922] first noted, d.f. = (s-1)(t-1))

While much work on two-dimensional contingency tables following the pioneering efforts by Pearson and Yule, it was not until 1935 that Bartlett, as a result of a suggestion by Fisher, utilized Yule's cross-product ratio to define the notion of second-order interaction in a 2×2×2 table, and to develop an appropriate test for the absence of such an interaction. The multivaraite generalizations of Bartlett's work, beginning with the work of Roy and Kastenbaum [1956], form the basis of the loglinear model approach to contingency tables, which is described in detail in Section 3.

The past 25 years has seen a burgeoning literature on the analysis of contingency tables, stemming in large part from work by S.N. Roy and his students at North Carolina, and from that of David Cox on binary regression. Some of this literature emphasizes the use of the minimum modified chi-square approach (e.g., Grizzle, Starmer, and Koch [1969]), or the use of the minimum discrimination information approach (e.g., Ku and Kullback [1968], and Gokhale and Kullback [1978]), while the bulk of it follows Fisher in the use of maximum likelihood. For most contingency table problems the minimum discrimination information approach yields maximum likelihood estimates.

Except for a few attempts at the use of additive models (see, e.g., Bhapkar and Koch [1968]) almost all the papers written on the topic emphasize the use of loglinear or logistic models. Key papers by Birch [1963], Darroch [1962], Good [1963], and Goodman [1963, 1964] plus the availability of high-speed computers, served to spur renewed interest in the problems of categorical data analysis. This in turn led to many articles by Leo Goodman (e.g., Goodman [1968, 1969, 1970]) and others, and finally culminated in books by Bishop, Fienberg and Holland [1975], Cox [1970], Gokhale and Kullback [1978], Haberman [1974], and Plackett [1974], all of which focus in large part on the use of loglinear models for both two-dimensional and multidimensional tables. A detailed bibliography for the statistical literature on contingency tables through 1974 is given by Killion and Zahn [1976].

The subsequent sections of this presentation are concerned primarily with the use of loglinear models for the analysis of contingency table data. For details on some related methods see the book by Lancaster [1969], and the series of papers on measures of association by Goodman and Kruskal, which have been recently reprinted as Goodman and Kruskal [1979]. Several book-length but elementary presentations on loglinear models are now available, including Everitt [1977], Fienberg [1980], Haberman [1978, 1979], and Upton [1978].

The next section describes two examples which will serve to illustrate some of the methods of analysis. Then, Section 3 briefly discusses some alternative methods for estimation of parameters used in conjunction with categorical data analysis, and Section 4 outlines the basic statistical theory associated with maximum likelihood estimation and loglinear models. These theoretical results are then illustrated, in Section 5 on the examples of Section 2. The final section concludes with a guide to (a) some recent applications of loglinear and contingency table modelling, and (b) computer programs for contingency table analysis.

2. Two Classic Examples

The data reported by Bartlett [1935] in his pioneering article, and included here in Table 1, are from an experiment giving the response (alive or dead) of 240 plants for each combination of the two explanatory variables, time of planting (early or late) and length of cutting (high or low).

Table 1: 2×2×2 table of Bartlett [1935]

2. Time of Planting		Early		Late	
3. Length of Cutting		High	Low	High	Low
1. Response	Alive	156	107	84	31
	Dead	84	133	156	209
Total		240	240	240	240

The questions to be answered are: (i) What are the effects of time of planting and length of cutting on survival? (ii) Do they interact in their effect on survival?

The data in Table 2, from Waite [1915], give the cross-classification or right-hand fingerprints according to the number of whorls and small loops. The total number of whorls and small loops is at most 5, and the resulting table is triangular:

Table 2: Fingerprints of the right hand classified by the number of whorls and small loops (Waite [1915])

	Small loops						
Whorls	0	1	2	3	4	5	Total
0	78	144	204	211	179	45	861
1	106	153	126	80	32		497
2	130	92	55	15			292
3	125	38	7				170
4	104	26					130
5	50						
Total	593	453	392	306	211	45	2000

Here the question of interest is more complicated because, as a result of the constraint forcing the data into the triangular structure, the number of whorls is "related to" the number of small loops. Such an array of counts is referred to as an *incomplete contingency table*, and the incomplete structure, in the case of the Waite data, was the source of yet another controversy involving Karl Pearson [1930], and, this time, J.A. Harris (see Harris and Treloar [1927]). In Section 5, the fit of a relatively simple model to these data is explored.

3. Estimating Parameters in Contingency Table Models

Let $\underset{\sim}{x}' = (x_1, x_2, \ldots, x_t)$ be a vector of observed counts for t cells, structured in the form of a cross-classification such as in Tables 1 and 2, where $t = 2^3 = 8$ and $t = 21$, respectively. Now let $\underset{\sim}{m}' = (m_1, m_2, \ldots, m_t)$ be the vector of expected values that are assumed to be functions of unknown parameters $\underset{\sim}{\theta}' = (\theta_1, \theta_2, \ldots, \theta_s)$, where $s < t$. Thus, one can write $\underset{\sim}{m} = \underset{\sim}{m}(\underset{\sim}{\theta})$.

There are three standard sampling models for the observed counts in contingency tables:

(i) Poisson model. The $\{x_i\}$ are observations from independent Poisson random variables with means $\{m_i\}$ and likelihood function:

$$\prod_{i=1}^{t} (m_i^{x_i} \exp(-m_i)/x_i!). \qquad (1)$$

(ii) Multinomial model. The total count $N = \sum_{i=1}^{t} x_i$ is a random sample from an infinite population where the underlying cell probabilities are $\{m_i/N\}$, and the likelihood is

$$N! \cdot N^{-N} \prod_{i=1}^{t} (m_i^{x_i}/x_i!). \qquad (2)$$

(iii) Product-Multinomial model. The cells are partitioned into sets, and each set has an independent multinomial structure, as in (ii).

For the Bartlett data in Section 2, the sampling model is product-multinomial -- there are actually 4 independent binomials, one for each of the 4 experimental conditions corresponding to the two factors time of planting and length of cutting. For the fingerprint data, the sampling model is multinomial. (See the discussion of factors and responses in the entry, Categorical Data, by Upton.)

For each of these sampling models the estimation problem can typically be structured in terms of a "distance" function, $K(\underset{\sim}{x}, \underset{\sim}{m})$, where parameter estimates $\hat{\underset{\sim}{\theta}}$ are chosen so that the distance between $\underset{\sim}{x}$ and $\underset{\sim}{m} = \underset{\sim}{m}(\underset{\sim}{\theta})$, as measured by $K(\underset{\sim}{x},\underset{\sim}{m})$, is minimized. The minimum chi-square method uses the distance function,

$$X^2(\underset{\sim}{x},\underset{\sim}{m}) = \sum_{i=1}^{t} (x_i-m_i)^2/m_i, \tag{3}$$

the minimum modified chi-square method uses the function

$$Y^2(\underset{\sim}{x},\underset{\sim}{m}) = \sum_{i=1}^{t} (x_i-m_i)^2/x_i, \tag{4}$$

and the minimum discrimination information method uses either

$$G^2(\underset{\sim}{x},\underset{\sim}{m}) = 2 \sum_{i=1}^{t} x_i \log (x_i/m_i), \tag{5}$$

or

$$G^2(\underset{\sim}{m},\underset{\sim}{x}) = 2 \sum_{i=1}^{t} m_i \log (m_i/x_i). \tag{6}$$

Rao [1962] studies these and other choices of "distance" functions.

For the three basic sampling models for contingency tables, choosing $\hat{\theta}$ to minimize $G^2(\underset{\sim}{x},\underset{\sim}{m})$ in (5) is equivalent to maximizing the likelihood function provided that

$$\sum_{i=1}^{t} m_i(\hat{\underset{\sim}{\theta}}) = \sum_{i=1}^{t} x_i, \tag{7}$$

(and that constraints similar to (7) hold for each of the set of cells under product-multinomial sampling, (iii)). Moreover, the estimators that minimize each of (3), (4), (5), and (6) in such circumstances belong to the class of Best Asymptotic Normal (BAN) estimates for $\underset{\sim}{m}$ (see Bishop, Fienberg, and Holland [1975] and Neyman [1949] for further discussion of asymptotic equivalence). Because of various additional asymptotic properties, and because of the smoothness of maximum likelihood estimates in relatively sparse tables,many authors have preferred to work with maximum likelihood estimates (MLE's), which minimize (5).

4. Some Basic Theory for Loglinear Models

For expected values $\{m_{ij}\}$ for a 2×2 table;

		B	
		1	2
A	1	m_{11}	m_{12}
	2	m_{21}	m_{22}

a standard measure of association for the row and column variables, A and B, respectively, is the cross-product ratio proposed by Yule [1900]:

$$\alpha = \frac{m_{11}m_{22}}{m_{12}m_{21}} \tag{8}$$

(for a discussion of the properties of α, see Bishop, Fienberg and Holland [1975] or Fienberg [1980]). Independence of A and B is equivalent to setting $\alpha = 1$, and can also be expressed in loglinear form:

$$\log m_{ij} = u + u_{1(i)} + u_{2(j)}, \tag{9}$$

where

$$\sum_{i=1}^{2} u_{1(i)} = \sum_{j=1}^{2} u_{2(j)} = 0. \tag{10}$$

Note that the choice of notation here parallels that for analysis of variance models. (See the entry, Categorical Data, by Upton for a related discussion, using somewhat different notation.)

Bartlett's [1935] no-second-order interaction model for the expected values in a 2×2×2 table

m_{111}	m_{121}
m_{211}	m_{221}

m_{112}	m_{122}
m_{212}	m_{222}

is based on equating the values of α in each layer of the table, i.e.,

$$\frac{m_{111}m_{221}}{m_{121}m_{211}} = \frac{m_{112}m_{222}}{m_{122}m_{212}} . \tag{11}$$

Expression (11) can be represented in loglinear form as

$$\log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)}, \tag{12}$$

where, as in (10), each subscripted u-term sums to zero over any subscript, e.g.,

$$\sum_i u_{12(ij)} = \sum_j u_{12(ij)} = 0. \tag{13}$$

All of the parameters in (12) can be written as functions of cross-product ratios (see Bishop, Fienberg, and Holland [1975]).

For the sampling schemes described in Section 3, the minimal sufficient statistics (MSS's) are the two-dimensional marginal totals, $\{x_{ij+}\}$, $\{x_{i+k}\}$, and $\{x_{+jk}\}$ (except for linearly redundant statistics included for purposes of symmetry), where a "+" indicates summation over the corresponding subscript. The MLE's of the $\{m_{ijk}\}$ under model (12) must satisfy the likelihood equations:

$$\begin{aligned} \hat{m}_{ij+} &= x_{ij+} && i,j = 1,2, \\ \hat{m}_{i+k} &= x_{i+k} && i,k = 1,2, \qquad (14) \\ \hat{m}_{+jk} &= x_{+jk} && j,k = 1,2, \end{aligned}$$

usually solved by some form of iterative procedure. For the Bartlett data the third set of equations in (14) corresponds to the binomial sampling constraints.

More generally, for a vector of expected values $\underset{\sim}{m}$, if the log-expectations $\underset{\sim}{\lambda}' = (\log m_1,\ldots,\log m_t)$ are representable as a linear combination of the parameters $\underset{\sim}{\theta}$, the following results hold under the Poisson and multinomial sampling schemes of Section 3:

(A) Corresponding to each parameter in $\underset{\sim}{\theta}$ is a MSS that is expressible as a linear combination of the $\{x_i\}$. (More formally, if $\mathcal{M}$ is used to denote the loglinear model specified by $\underset{\sim}{m} = \underset{\sim}{m}(\underset{\sim}{\theta})$, then the MSS's are given by the projection of $\underset{\sim}{x}$ onto $\mathcal{M}$, $P_{\mathcal{M}}\underset{\sim}{x}$. For a more detailed discussion see Haberman [1974].)

(B) The MLE, $\hat{\underset{\sim}{m}}$, of $\underset{\sim}{m}$, if it exists, is unique and satisfies the likelihood equations:

$$P_{\mathcal{M}}\hat{\underset{\sim}{m}} = P_{\mathcal{M}}\underset{\sim}{x}. \qquad (15)$$

(Note that the equations in (14) are a special case of those given by expression (15).)

Necessary and sufficient conditions for the existence of a solution to the likelihood equations, (15), are relatively complex (see Haberman [1974]). A sufficient condition is that all cell counts be positive, i.e., $\underset{\sim}{x} > 0$, but MLE's for loglinear models exist in many sparse situations where a large fraction of the cells have zero counts.

For product-multinomial sampling situations, the basic multinomial constraints (i.e., that the counts must add up to the multinomial sample sizes) must be taken into account. Typically, some of the parameters in $\underset{\sim}{\theta}$ which specify the loglinear model $\mathcal{M}$, i.e., $\underset{\sim}{m} = \underset{\sim}{m}(\underset{\sim}{\theta})$, are fixed by these constraints.

More formally, let $\mathcal{M}^*$ be a loglinear model for $\underset{\sim}{m}$ under product-multinomial sampling which corresponds to a loglinear model $\mathcal{M}$ under Poisson sampling, such that the multinomial constraints "fix" a subset of the parameters, $\underset{\sim}{\theta}$, used to specify $\mathcal{M}$. Then

(C) The MLE of $\underset{\sim}{m}$ under product-multinomial sampling for the model $\mathcal{M}^*$ is the same as the MLE of $\underset{\sim}{m}$ under Poisson sampling for the model $\mathcal{M}$.

As a consequence of Result C, equations (14) are the likelihood equations for the 2×2×2 table under the no-second-order interaction model for Poisson or multinomial sampling, as well as for product-multinomial sampling when any set of one-way or two-way marginal totals are fixed (i.e., these correspond to the multinomial constraints).

A final result, that is used to assess the fit of loglinear models, can be stated in the following informal manner:

(D) If $\hat{\underset{\sim}{m}}$ is the MLE of $\underset{\sim}{m}$ under a loglinear model, and if the model is correct, then the statistics

$$X^2 = \sum_{i=1}^{t} (x_i - \hat{m}_i)^2 / \hat{m}_i \qquad (16)$$

and

$$G^2 = 2 \sum_{i=1}^{t} x_i \log (x_i / \hat{m}_i) \qquad (17)$$

have asymptotic χ^2 distributions with t-s degrees of freedom, where s is the total number of independent constraints implied by the loglinear model and the multinomial sampling constraints (if any). If the model is not correct then X^2 and G^2, in (16) and (17), are stochastically larger than χ^2_{t-s}. (See the entry, Chi-square Tests, by Bhapkar and Koch.) Expression (17) is the minimizing value of the distance function (5), but (16) is not the minimizing chi-square value for the function (3).

In the next section these basic results are applied in the context of the Bartlett and Waite data sets of Section 2.

Many authors have devised techniques for selecting among the class of loglinear models applicable foı contingency table structures. These typically (although not always) resemble corresponding model selection procedures for analysis of variance and regression models. See, for example, Goodman [1971] and Aitken [1978], as well as the discussions in Bishop, Fienberg, and Holland [1975], and Fienberg [1980].

5. Contingency Table Analyses

5.1 Illustrative Analyses

For the 2^3 table of Bartlett from Section 2, variables 2 and 3 are fixed by design, so that $\hat{m}_{+jk} = 240$, and the estimated expected values under the no second-order interaction model of expression (12) are given in Table 3. These values were computed by Bishop, Fienberg and Holland [1975] using the method of iterative proportional fitting. Bartlett originally found the solution to equations (14), by noting that the constraints in his specification, (11), reduced (14) to a single cubic equation for the discrepancy $\Delta = \hat{m}_{111} - x_{111}$. Note that the expected values satisfy expression (12), e.g., $\hat{m}_{12+} = 78.9 + 36.1 = 115 = 84 + 31 = x_{12+}$. The goodness-of-fit statistics for this model are $X^2 = 2.27$ and $G^2 = 2.29$. Using Result D of Section 4, one compares these values to tail-values of the chi-square distribution with 1 d.f., e.g. $\chi_1^2(.10) = 2.71$, and this suggests that the no-second-order interaction model provides an acceptable fit to the data.

Since the parameters u, $\{u_{2(j)}\}$, $\{u_{3(k)}\}$ and $\{u_{23(jk)}\}$ are fixed by the binomial sampling constraints for these data, model (12) is often rewritten as

$$\log \left(\frac{m_{1jk}}{m_{2jk}}\right) = 2[u_{1(1)} + u_{12(1j)} + u_{13(2k)}]$$
$$= w + w_{2(j)} + w_{3(k)}, \qquad (18)$$

where

Table 3: Observed and Expected Values for the Bartlett Data Including the No-Second Order Interaction Model.

Cell	Observed x	Estimated Expected $\hat{m}$
1,1,1	156	161.1
2,1,1	84	78.9
1,2,1	84	78.9
2,2,1	156	161.1
1,1,2	107	101.9
2,1,2	133	138.1
1,2,2	31	36.1
2,2,2	209	203.9

$$\sum_j w_{2(j)} = \sum_k w_{3(k)} = 0.$$

Expression (18) is referred to as a logit model for the log-odds for alive versus dead. The simple additive structure corresponds to Bartlett's notion of no second-order interaction.

For the Waite fingerprint data of Table 2, one model that has been considered is the simple additive loglinear model of expression (9), but only for those cells where positive counts are possible, i.e., in the upper triangular section. For cells with $i > j$, $m_{ij} = 0$ a priori. This restricted version of the independence model is referred to as quasi-independence, and the results of the preceding section can be used in connection with it. The MSS's are still the row and column totals (Result A). The likelihood equations under multinomial sampling are (applying Results B and C):

$$\hat{m}_{i+} = x_{i+} \qquad i = 0,1,2,\ldots,5$$
$$\hat{m}_{+j} = x_{+j} \qquad j = 0,1,2,\ldots,5, \tag{20}$$

where $m_{ij} = 0$ for $i > j$. A solution of equations (20) satisfying the model can be found directly (see Goodman [1968] or Bishop and Fienberg [1969]), or by using a standard iterative procedure. The estimated expected values for the fingerprint data under the model of quasi-independence are given in Table 4, and they satisfy the marginal constraints in expression (20).

Table 4. Estimated Expected Values for Fingerprint Data Under Quasi-Independence

	Small loops						
Whorls	0	1	2	3	4	5	Total
0	200.6	167.4	166.6	150.3	131.1	45.0	861
1	122.2	101.9	101.4	91.6	79.9		497
2	85.5	71.4	71.0	64.1			292
3	63.8	53.2	53.0				170
4	70.9	59.1					130
5	50.0						50
Total	593	453	392	306	211	45	2000

The goodness-of-fit statistics for this model are $X^2 = 399.8$ and $G^2 = 450.4$ which correspond to values in the very extreme right-hand tail of the χ^2_{10} distribution. Thus the model of quasi-independence seems inappropriate. Darroch [1971] describes the loglinear model of F-independence (with more parameters than the quasi-independence model), which takes in account the way in which the constraint, that the number of small loops plus the number of whorls cannot exceed 5, makes the usual definition of independence inappropriate. This model in loglinear form is

$$\log \hat{m}_{ij} = u + u_{1(i)} + u_{2(j)} + u_{3(5-i-j)}, \qquad (21)$$

where the u_3-parameters correspond to diagonals where the sum of the numbers of whorls and small loops is constant. Darroch and Ratcliff [1973] illustrate the fit of the F-independence model to a related set of fingerprint data involving large rather than small loops.

5.2 Multidimensional Contingency Table Analysis

Not all applications of loglinear models involve such simple structures as 2^3 tables, or even incomplete 6×6 arrays. Indeed, much of the methodology was developed in the mid-1960's to deal with very large, highly multidimensional tables. For example, in the National Halothane Study (Bunker et al. [1969]), investigators considered data on the use of (i) 5 anesthetic agents, in operations involving (ii) 4 levels of risk, and patients of (iii) 2 sexes, (iv) 10 age groups, with (v) 7 differing physical statuses (levels of anesthetic risk) and (vi) previous operations (yes, no), for (vii) 3 different years, from (viii) 34 different institutions. Two sets of data were collected, the first consisting of all deaths within six weeks of surgery, and the second consisting of a sample (of comparable size) of all those exposed to surgery. Thus the data consisted of two very sparse 5×4×2×10×7×2×3×34 tables, each containing in excess of 57,000 cells. One of the more successful approach used in the analysis of the data in these tables was based on loglinear models and the generalizations of the methods illustrated in this section.

One of the key reasons why loglinear models have become so popular in such analyses is that they lead to a simplified description of the data in terms of marginal totals -- the minimal sufficient statistics of Result A of Section 4. This is especially important when the table of data is large and sparse. For more details on the Halothane Study analyses, as well as examples of other applications involving four-way and higher dimensional tables of counts, see Bishop, Fienberg, and Holland [1975].

A second reason for the popularity of loglinear models relates to their interpretation. A large subset of these models can be interpretted in terms of independence or the conditional independence of several discrete random variables given the values of other discrete variables, thus generalizing the simple ideas for 2×2 tables outlined in Section 4. For further details, see any of the books cited in Section 1.

6. A Brief Guide to Additional Applications and Computing Programs

6.1 Novel Applications Involving Contingency Tables

Many data sets can profitably be structured to appear in the form of a cross-classification of counts, and then analyzed using methods related to those described in this entry. Some examples of applications where this has been done include the following:

(a) Capture-multiple-recapture analysis to estimate the size of a non-changing population (Fienberg [1972], Bishop, Fienberg and Holland [1975]). If the members of non-changing populations are sampled k successive times (possibly dependent), then the resulting recapture history data can be displayed in the form of a 2^k table with one missing cell, corresponding to those never sampled. Such an array is amenable to log-linear analysis, the results of which can be used to project a value for the missing cell.

(b) Guttman scaling of a sequence of p dichotomous items (Goodman [1975]). The items form a perfect Guttman scale if they have an order such that a positive response to any item implies a positive response to those items lower in the ordering. Goodman describes an application of techniques for incomplete multidimensional contingency tables in which he measures departures from perfect Guttman scales.

(c) Latent structure analysis, where unobservable categorical variables are included as part of the analysis of categorical data structures, and the observable variables are taken to be conditionally independent given the unobservable latent variables (Goodman [1974]; see also the entry, Categorical Data, by Upton).

(d) Paired comparisons of several objects by a set of judges, with the outcome being the preference of one object over the other. A well-known model for paired comparisons, first proposed by Bradley and Terry, and several extensions to it, can be viewed as loglinear models. Then relatively standard contingency table methods can be used to analyze pair comparisons data (see Imrey, Johnson, and Koch [1976], Fienberg and Larntz [1976], and Fienberg [1979]).

6.2 Computer Programs for Loglinear Model Analysis

As with other forms of multivariate analysis, the analysis of multidimensional contingency tables relies heavily on computer programs. A large number of these have been written to compute estimated parameter values for loglinear models and associated test statistics, and most computer installations at major universities have one or more programs available for users.

The most widely used numerical procedure for the calculation of maximum likelihood estimates for loglinear models is the method of iterative proportional fitting (IPF), which iteratively adjusts the entries of a contingency table to have marginal totals equal to those used in specifying the likelihood equations. Detailed Fortran listings for this method are available in Haberman [1972, 1973], and they have been implemented in the BMDP Programs distributed by the UCLA Health Sciences Computing Facility (Dixon and Brown [1979]), as well as in a variety of other forms. IPF programs also exist in other languages such as APL (e.g., see Fox [1979]). The major advantage of the IPF method is that it requires limited computer memory capabilities since it does not require matrix inversion or equivalent computations, and thus can be used in connection with the analysis of very high dimensional tables. Its major disadvantage is that it does not provide, in an easily accessible form, estimates of the basic loglinear model parameters (and an estimate of their asymptotic covariance matrix); it only provides estimated expected values.

The other numerical approaches suggested for the computation of maximum likelihood estimates are typically based on classical procedures for solving nonlinear equations such as modifications of Newton's method or the Newton-Raphson method (e.g., see the listing in Haberman [1979]). Currently the most widely used such program is GLIM, distributed by the Numerical Algorithms Group of the United Kingdom (Baker and Nelder [1978]), which fits a class of generalized linear models, of which loglinear and logit models are special cases. The virtue of these programs is that they produce both estimated expected values, and estimated parameter values and an estimate of the asymptotic covariance matrix. Unfortunately, such output comes at the expense of added storage and these programs cannot handle analyses for very large contingency tables. Several groups of researchers are currently at work adapting variants of Newton's method using numerical techniques that will allow for increased storage capacity, and thus the analysis of larger tables than is currently possible.

Computation problems remain as a major stumbling block to the widespread application of loglinear model methods to the analysis of large data sets structured in the form of multidimensional cross-classifications of counts.

7. Bibliography

7.1 Books on the Analysis of Contingency Tables

Bishop, Y.M.M., Fienberg, S.E., and Holland, P. [1975]. Discrete Multivariate Analysis: Theory and Practice. Cambridge, Mass.: The MIT Press.

[A systematic exposition and development of the loglinear model for the analysis of contingency tables, primarily using maximum likelihood estimation, and focussing on the use of iterative proportional fitting. Includes chapters on measures of association, and others on special related topics. Contains both theory and numerous examples from many disciplines with detailed analyses.]

Cox, D.R. [1970]. Analysis of Binary Data. London: Methuen.

[A concise treatment of loglinear and logistic response models, primarily for binary data. Emphasis on statistical theory, especially related to exact tests; includes several examples.]

Everitt, B.S. [1977]. The Analysis of Contingency Tables. London: Chapman and Hall.

[A brief and very elementary introduction to contingency table analysis, with primary emphasis on two-dimensional tables.]

Fienberg, S.E. [1980]. The Analysis of Cross-Classified Categorical Data (2nd ed.). Cambridge, Mass.: The MIT Press.

[A comprehensive introduction, for those with some training in statistical methodology, to the analysis of categorical data using loglinear models and maximum likelihood estimation. Emphasis on methodology, with numerous examples and problems.]

Gokhale, D.V. and Kullback, S. [1978]. The Information in Contingency Tables. New York: Marcel Dekker.

[A development of minimum discrimination information procedures for linear and loglinear models. Contains a succinct theoretical presentation, followed by numerous examples.]

Goodman, L.A. and Kruskal, Wm. [1979]. Measures of Association for Cross Classifications. New York: Springer-Verlag.

[A reprinting of four classical papers, written between 1954 and 1972, on the construction of measures of association for two-way tables, historical references, sample estimates,and related asymptotic calculations.]

Haberman, S.J. [1974]. The Analysis of Frequency Data. Chicago: University of Chicago Press.

[A highly mathematical,advanced presentation of statistical theory associated with loglinear models and of related statistical and computational methods. Contains examples, but is suitable only for mathematical statisticians who are familiar with the topic.]

Haberman, S.J. [1978]. Analysis of Qualitative Data, Volume 1 (Introductory Topics). New York: Academic Press.

Haberman, S.J. [1979]. Analysis of Qualitative Data, Volume 2 (New Developments). New York: Academic Press.

[An intermediate level, two-volume introduction to the analysis of categorical data via loglinear models, emphasizing maximum likelihood estimates computed via the Newton-Raphson algorithm. Volume 1 examines complete cross-classifications, and Volume 2 considers multinomial response models, incomplete tables, and other related topics. Contains many examples, problems and solutions, and a computer program listing (for two-way tables) in Volume 2.]

Lancaster, H.O. [1969]. The Chi-Squared Distribution, Chapters 11 and 12. New York: John Wiley.

[A mathematical statistics monograph developing ideas on the chi-square distribution and quadratic forms for both discrete and continuous random variables, with several chapters related to the analysis of contingency tables. Emphasis is on topics other than loglinear models.]

Plackett, R.L. [1974]. The Analysis of Categorical Data. London: Griffin.

[A concise introduction to statistical theory and methods for the analysis of categorical data. Assumes a thorough grasp of basic principles of statistical inference. Considerable emphasis on two-way tables. Contains many examples and exercises.]

Upton, G.J.G. [1978]. The Analysis of Cross-Tabulated Data. New York: John Wiley.

[A brief introduction to the analysis of contingency tables via loglinear models and measures of association for those with some training in statistical methodology. Contains several examples.]

7.2 Computer Program Descriptions and Documentation

Baker, R.J. and Nelder, J.A. [1978]. The GLIM System, Release 3, Manual. Oxford, England: Numerical Algorithms Group.

Bock, R.D. and Yates, G. [1973]. MULTIQUAL: Log-linear Analysis of Nominal or Ordinal Qualitative Data by the Method of Maximum Likelihood. Chicago, Ill.: International Education Services.

[A manual for a loglinear model program that uses a modified Newton-Raphson algorithm.]

Dixon, W.J. and Brown, M.B. (eds.) [1979]. BMPD, Biomedical Computer Programs, P-Series. Berkeley, Ca.: University of California Press.

[See Chapter 11 on frequency tables and Section 14.LR on logistic regression.]

Fox, J. [1979]. "TAB: An APL Workspace for the Log-linear Analysis of Contingency Tables." American Statistician, 33, 159-160.

[Contains a program description, but no listing.]

Haberman, S.J. [1972]. "Loglinear Fit for Contingency Tables (Algorithm AS 51)." Applied Statistics, 21, 218-225.

[Contains Fortran listing of program that uses iterative proportional fitting.]

Haberman, S.J. [1973]. "Printing Multidimensional Tables (Algorithm AS 57)." Applied Statistics, 22, 118-126.

[Contains Fortran listing of program.]

SAS Institute [1979]. SAS User's Guide (1979 Edition). Raleigh, N.C.: SAS Institute.

[See pages 298-301 for instructions on the use of general programs for nonlinear equations for computing minimum modified chi-square estimates, and maximum likelihood estimations using a Newton-Raphson algorithm.]

7.3 Other References Cited in Entry

Aitken, M. [1978]. "The Analysis of Unbalanced Cross-Classifications (with Discussion)." Journal of the Royal Statistical Society, Series A, 141, 195-223.

Bartlett, M.S. [1935]. "Contingency Table Interactions." Journal of the Royal Statistical Society, Supplement, 2, 248-252.

Bhapkar, V.P. and Koch, G. [1968]. "On The Hypotheses of 'No Interaction' in Contingency Tables." Biometrics, 24, 567-594.

Birch, M.W. [1963]. "Maximum Likelihood in Three-Way Contingency Tables." Journal of the Royal Statistical Society, Series B, 25, 229-233.

Bishop, Y.M.M. and Fienberg [1969]. "Incomplete Two-Dimensional Contingency Tables." Biometrics, 25, 119-128.

Bunker, J.P., Forrest, W.H., Jr., Mosteller, F., and Vandam L. [1969]. The National Halothane Study. Report of the Subcommittee on the National Halothane Study of the Committee on Anesthesia, Division of Medical Sciences, National Academy of Sciences-National Research Council, National Institutes of Health, National Institute of General Medical Sciences, Bethesda, Maryland. Washington, D.C.: U.S. Government Printing Office.

Darroch, J.N. [1962]. "Interaction in Multi-Factor Contingency Tables." Journal of the Royal Statistical Society, Series B, 24, 251-263.

Darroch, J.N. [1971]. "A Definition of Independence for Bounded-Sum, Nonnegative, Inter-Valued Variables." Biometrika, 58, 357-368.

Darroch, J.N. and Ratcliff, D. [1973]. "Tests of F-Independence with Reference to Quasi-Independence of Waites's Fingerprint Data." Biometrika, 60, 395-402.

Fienberg, S.E. [1972]. "The Multiple-Recapture Census for Closed Populations and Incomplete 2^k Contingency Tables." Biometrika, 59, 591-603.

Fienberg, S.E. [1979]. "Loglinear Representation for Paired Comparison Models with Ties and Within-Pair Order Effects." Biometrics, 35, 479-481.

Fienberg, S.E. and Larntz, K. [1976]. "Loglinear Representation for Paired and Multiple Comparisons Models." Biometrika, 63, 245-254.

Fisher, R.A. [1922]. "On the Interpretation of χ^2 from Contingency Tables, and the Calculation of P." Journal of the Royal Statistical Society, 85, 87-94.

Good, I.J. [1963]. "Maximum Entropy for Hypotheses Formulation Especially for Multidimensional Contingency Tables." The Annals of Mathematical Statistics, 34, 911-934.

Goodman, L.A. [1963]. "On Methods for Comparing Contingency Tables." Journal of the Royal Statistical Society, Series A, 126, 94-108.

Goodman, L.A. [1964]. "Simultaneous Confidence Limits for Cross-Product Ratios in Contingency Tables." Journal of the Royal Statistical Society, Series B, 26, 86-102.

Goodman, L.A. [1968]. "The Analysis of Cross-Classified Data: Independence, Quasi-Independence, and Interaction in Contingency Tables With or Without Missing Cells." Journal of the American Statistical Association, 63, 1091-1131.

Goodman, L.A. [1969]. "On Partitioning X^2 and Detecting Partial Association in Three-Way Contingency Tables." Journal of the Royal Statistical Society, Series B, 31, 486-498.

Goodman, L.A. [1971]. "The Analysis of Multidimensional Contingency Tables: Stepwise Procedures and Direct Estimation Methods for Building Models for Multiple Classifications." Technometrics, 13, 33-61.

Goodman, L.A. [1974]. "Exploratory Latent Structure Analysis Using Both Identifiable and Unidentifiable Models." Biometrika, 61, 215-231.

Goodman, L.A. [1975]. "A New Model for Scaling Response Patterns: An Application of the Quasi-Independence Concept." Journal of the American Statistical Association, 70, 755-768.

Grizzle, J.E., Starmer, C.F., and Koch, G.G. [1969]. "Analysis of Categorical Data by Linear Models." Biometrics, 25, 489-504.

Harris, J.A. and Treloar, A.E. [1927]. "On a Limitation in the Applicability of the Contingency Coefficient." Journal of the American Statistical Association, 22, 460-472.

Imrey, P.B., Johnson, W.D., and Koch, G.G. [1976]. "An Incomplete Table Approach to Paired-Comparison Experiments." Journal of the American Statistical Association, 71, 614-623.

Killion, R.A. and Zahn, D.A. [1976]. "A Bibliography of Contingency Table Literature: 1900-1974." International Statistical Review, 44, 71-112.

Ku, H.H. and Kullback, S. [1968]. "Interaction in Multidimensional Contingency Tables: An Information Theoretic Approach." Journal of Research of the National Bureau of Standards, 72B, 159-199.

Neyman, J. [1949]. "Contributions to the Theory of the χ^2-Test." Proceedings of Berkeley Symposium on Mathematical Statistics and Probability. Berkeley: University of California Press, 239-273.

Pearson, K. [1900a]. "On a Criterion that a Given System of Deviations from the Probable in the Case of a Correlated System of Variables is such that it can be Reasonably Supposed to have Arisen from Random Sampling." Philosphical Magazine and Journal of Science, 50, 157-175.

Pearson, K. [1900b]. "Mathematical Contributions to the Theory of Evolution in the Inheritance of Characters Not Capable of Exact Quantitative Measurements, VIII." Philosophical Transactions of the Royal Society of London, Series A, 195, 79-150.

Pearson, K. [1904]. "Mathematical Contributions to the Theory of Evolution. XIII. On the Theory of Contingency and Its Relation to Association and Normal Correlation." Draper's Company Research Memoirs, Biometric Series I, 1-35.

Pearson, K. [1930]. "On the Theory of Contingency. I. Note on Professor J. Arthur Harris' Paper on the Limitations in the Applicability of the Contingency Coefficient." Journal of the American Statistical Association, 25, 320-323.

Pearson, K. and Heron, D. [1913]. "On Theories of Association." Biometrika, 9, 159-315.

Quetelet, M.A. [1849]. Letters Addressed to H.R.H. the Grand Duke of Saxe Coburg and Gotha on the Theory of Probabilities as Applied to the Moral and Political Sciences (translated from the French by Olinthus Gregory Downs). London, England: Charles and Edwin Layton.

Rao, C.N. [1962]. "Efficient Estimates and Optimum Inference Procedures in Large Samples (with Discussion)." Journal of The Royal Statistical Society, Series B, 24, 46-72.

Roy, S.N. and Kastenbaum, M.A. [1956]. "On the Hypothesis on No 'Interaction' in a Multi-Way Contingency Tables." The Annals of Mathematical Statistics, 27, 749-757.

Waite, H. [1915]. "Association of Finger-Prints." Biometrika, 421-478.

Yule, G.U. [1900]. "On the Association of Attributes in Statistics: With Illustration from the Material of the Childhood Society." Philosophical Transactions of the Royal Society of London, Series A, 194, 257-319.

8. Acknowledgment

The preparation of this entry was supported in part by Contract N00014-7-C-0600 from the Office of Naval Research, Statistics and Probability Program, to the School of Statistics, University of Minnesota.

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER
(14) TR-
Technical Report No. 369

2. GOVT ACCESSION NO.
AD-A084262

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
(6) Contingency Tables

5. TYPE OF REPORT & PERIOD COVERED

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)
(10) Stephen E. Fienberg

8. CONTRACT OR GRANT NUMBER(s)
(15) NØØ014-78-C-Ø600

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Department of Applied Statistics
School of Statistics, University of Minnesota
1994 Buford Avenue, St. Paul, MN 55108

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS

11. CONTROLLING OFFICE NAME AND ADDRESS
Office of Naval Research
800 N. Quincy Street
Arlington, VA 22217

12. REPORT DATE
(11) Feb[redacted] 1[redacted]80

13. NUMBER OF PAGES
30

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)
(12) 32

15. SECURITY CLASS. (of this report)
Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)
Categorical data; logit models; loglinear models; iterative proportional fitting

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This paper is concerned primarily with the use of loglinear models for the analysis of contingency table data. An introduction to multivariate statistical analysis is presented and the topic is placed in an historical perspective. Examples which illustrate some of the methods of analysis are described. The basic statistical theory associated with maximum likelihood estimation and loglinear models is outlined and the results are applied to the examples. The paper concludes with a guide to some recent applications of loglinear and contingency table modelling, and to computer programs for contingency table analysis.

411265

13